Automatic Example-based Image Colourisation using Location-Aware Cross-Scale Matching

Bo Li, Yu-Kun Lai, Matthew John, Paul L. Rosin

***Abstract*—Given a reference colour image and a destination grayscale image, this paper presents a novel automatic colourisation algorithm that transfers colour information from the reference image to the destination image. Since the reference and destination images may contain content at different or even varying scales (due to changes of distance between objects and the camera), existing texture matching based methods can often perform poorly. We propose a novel cross-scale texture matching method to improve the robustness and quality of the colourisation results. Suitable matching scales are considered locally, which are then fused using global optimisation that minimises both the matching errors and spatial change of scales. The minimisation is efficiently solved using a multi-label graph-cut algorithm. Since only low-level texture features are used, texture matching based colourisation can still produce semantically incorrect results, such as meadow appearing above the sky. We consider a class of semantic violation where the statistics of up-down relationships learnt from the reference image are violated and propose an effective method to identify and correct unreasonable colourisation. Finally, a novel nonlocal ℓ_1 optimisation framework is developed to propagate high confidence micro-scribbles to regions of lower confidence to produce a fully colourised image. Qualitative and quantitative evaluations show that our method outperforms several state-of-the-art methods.**

***Index Terms*—Image colourisation, cross-scale texture matching, location statistics, graph cut, sparse, edge preserving**

I. Introduction

Image colourisation is the process of adding colour to grayscale images. The uses for colourisation of grayscale images are numerous ranging from converting black and white movies to colour, to colourising historic photographs to improve the aesthetics of the image. The alternative problem, conversion of colour images to grayscale, is generally straightforward although some methods [1], [2] better differentiate the chromatic difference of pixels. A popular process of image colourisation is one whereby the user scribbles a few strokes of colour on the image, and the colours of the remaining pixels are then determined automatically [3]. The results of the colourisation can differ greatly depending upon how the colour scribbles are chosen, hence results depend upon user skill and experience. This is exemplified when a novice user applies the colour scribbles forgetting to mark the region boundaries where intensities are similar, causing the colourisation algorithm to spread the colour to regions of the image that should not contain the particular colour, producing unrealistic results.

Bo Li is with the School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China, and also with the School of Educational Information Technology, Central China Normal University, Wuhan, China. e-mail: bolimath@gmail.com.

Yu-Kun Lai, Matthew John and Paul Rosin are with the School of Computer Sciences and Informatics, Cardiff University, Cardiff, UK.

Automatic colourisation methods such as that by Welsh et al. [4] have removed the burden of annotating the image with colour scribbles by using a colour reference image to transfer colour. They are able to produce realistic results if a suitable reference colour image is provided by the user. Such work focuses on performing colour transfer automatically from reference images and propagating the colour from a small number of transferred colour scribbles. In contrast with the task of traditional colour transfer, the destination image does not have colour information, so colourisation mainly relies on matching of luminance and texture information. In many cases, the objects have different scales in the reference and destination images. Therefore, feature matching in different scales is essential. Although some advanced feature detectors, such as SIFT [5] and SURF [6], can provide scale-invariant features, they are sparse and not suitable for dense matching required for colourisation. In this paper, we propose a *cross-scale* matching method that considers different potential scales locally when matching the reference and destination images, which are then fused globally with graph-cut to find spatially coherent scales with good matching quality. Colour transfer via texture matching may result in some semantic errors, e.g. when similar textures cause confusion, some of the sky may be colourised in the colour of grass. Such unreasonable colourisation results cannot be detected by low-level texture features, although appear obviously wrong to a human observer. Instead of using machine learning which requires a large number of training images and loses the flexibility of easily specifying the desired colour style, we focus on a class of simple semantic violations where the up-down spatial relationship is violated. This is a reasonable assumption as images are normally taken with an (near) upright camera. We perform statistics of up-down relationships of colour distribution in the reference image, which is then used to help detect and correct unreasonable matching results in the destination image, as we assume the content of the reference and destination images are semantically related. Finally, a colour propagation step is performed to diffuse colour from a small number of scribbles to the whole image. Improper propagation may cause over-smoothing effect at edges or step effect in flat regions. In this paper, we propose a nonlocal ℓ_1 optimisation framework along with confidence weighting to suppress artefacts caused by wrong matchings while avoiding over-smoothing edges.

Examples of our colourisation method are shown in Fig. 1 where the same destination image is colourised with different reference images to obtain different but all plausible results. Such flexibility is often necessary to avoid essential semantic and artistic ambiguities. Popular deep-learning methods can learn automatic colour mapping through training on a large

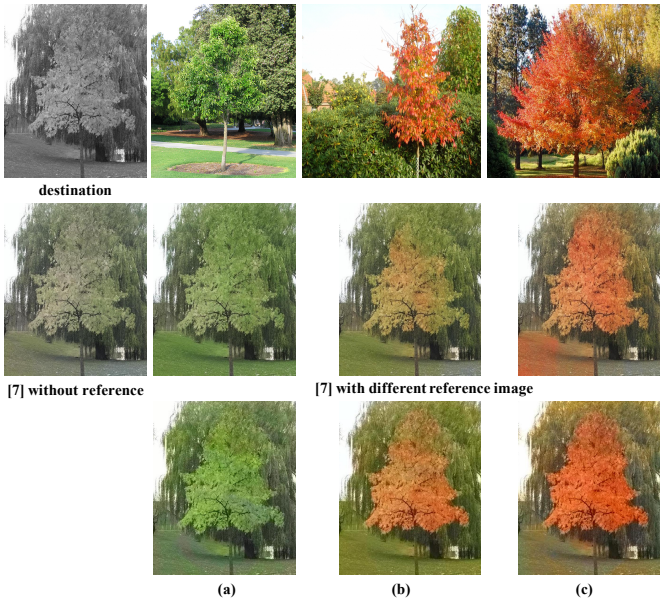

Fig. 1. Colourisation results with different reference images. The first row is the destination grayscale image and different reference images, the second row and the third row are the corresponding colourisation results of method [7] and the proposed method respectively.

set of images, and have shown the powerful ability for the task of image colourisation. However, most of the existing deep-learning colourisation methods cannot produce example-based colourisation results. Although the most recent state-of-the-art deep learning based method [7] can also cooperate with a reference image, its mechanism is a global colour distribution mapping, which can result in error colours for challenging cases, as shown in the second row. More examples of method [7] will be shown in Sec. IV.

The major contributions of this paper are:

1) To address scale variation within and between reference and destination images, we propose a cross-scale matching method, where locally good matching scales are identified, which are then fused using a global graph-cut based optimisation. This approach improves robustness and quality of colourisation.
2) We consider a class of semantic violation typically appearing in automatic colourisation and develop an effective solution to automatically detect and correct unreasonable matching results, using statistics of up-down colour distribution in the reference images. Our method significantly improves colourisation results for challenging cases where existing texture matching based methods fail.
3) Instead of using a small set of user scribbles, we propose a novel confidence weighted nonlocal ℓ_1 colour propagation method in which a dense micro-scribble image composed of matched colours is generated along with a confidence map used as a soft constraint in the optimisation framework. The proposed nonlocal ℓ_1 propagation framework maintains the tone of micro-scribble image while effectively suppresses the over-smoothing effect at edges.

In the following sections, we first review related work in Sec. II. We describe our method in detail in Sec. III, followed by experimental results both qualitatively and quantitatively in Sec. IV. Finally conclusions are drawn in Sec. V.

II. Related work

Some work considers interactive colourisation guided by user scribbles. Levin et al. [3] presented a colourisation method based upon an optimisation problem. The user has to apply several colour scribbles to an image and the colours are then propagated through the image by means of minimising a quadratic cost function. The optimisation problem is formulated as a linear system which can be solved efficiently, resulting in a colourised image in a relatively moderate timescale. Yatziv and Sapiro [8] proposed another scribble based colourisation algorithm. It is based on luminance-weighted chrominance blending and efficient intrinsic distance computation, and leads to a more efficient algorithm for both image and video colourisation. Nie et al. [9] proposed another improvement over [3] with comparable quality and improved efficiency by using quadtree decomposition based non-uniform sampling. Although scribble based methods can colourise images to a high standard, it requires a significant amount of user effort to apply the colour scribbles to the grayscale image. Examples given by the authors show a varying number of scribbles required dependent on the complexity of the image. Balinsky and Mohammad [10] put forward a Bayesian analysis of the colourisation problem that is convexified by using ℓ_1 optimisation. Similar to [3] the image is annotated with colour scribbles. The authors show that their proposed approach outperforms that of [3] although their method takes longer to solve an equivalent sized image. For all these scribble based methods, the colourisation results are highly influenced by the placement of colour scribbles, thus to achieve realistic results the user has to have a degree of knowledge regarding the effect of scribble placement, and it can be time-consuming.

Welsh et al. [4] proposed the first automatic colourisation method using a colour transfer approach from a reference colour image to a destination grayscale image. The method transfers the chromatic information from the reference image to the destination image based upon local matches of the weighted average of a pixel's luminance and neighbourhood statistics keeping the luminance of the destination image unchanged. Such colour transfer works well when there is a strong correlation between the luminance values of colour regions in the reference and destination images. For more challenging images, the idea of swatches was introduced whereby user interaction was incorporated to indicate regions in the reference image that should be transferred to the destination image. This work was extended by Ying et al. [11] by using a more extensive neighbourhood description computed using co-occurrence matrix based texture features. To reduce the artefacts caused by outliers, the edit-nearest-neighbour method [12] is used to try to remove the outliers. While producing improved results, the co-occurrence matrix is expensive to compute. Chen et al. [13] combined [4] with foreground/background image matting to improve the

colourisation results but user interaction is needed to guide the grayscale image matting. A global histogram regression based method is proposed in [14]. The method is based on the assumption that the colourised image should have similar colour distribution as the reference. The method however may not produce ideal results for complicated images where global histogram mapping is not sufficient.

A couple of variational models were proposed for image colourisation in [15], which appropriately added edge information from the brightness data, while reconstructing smooth colour values for each homogeneous region. To improve spatial consistency and suppress colour bleeding, Bugeau et al. [16] proposed an image colourisation method based on an edge-preserving total variation formulation, which only involves chrominance channels. As a result, the method may produce halo effects near strong edges. To circumvent this, Pierre et al. [17] proposed an improved method with a regularisation involving both luminance and chrominance information, which helps to better preserve edge structures in the colourised images. To improve texture matching, especially near edges, Arbelot et al. [18] developed a colour transfer and colourisation method that utilises spatial coherence around image structure by adopting an edge-aware texture descriptor based on region covariance, although local matching is still performed independently. A locality consistent sparse representation learning method is proposed by [19]. By incorporating locality consistency in the matching stage rather than in post-processing as existing methods do, it substantially improves colour consistency and reduces artefacts.

Irony et al. [20] developed a colourisation method that takes account of the context of pixels rather than attempting to colourise a pixel based upon its neighbourhood statistics alone. The approach first segments the reference colour image by using a supervised classification scheme, then a mapping is made between small neighbourhood areas and points in feature space. The mapping discriminates between pixels which come from different image regions based upon the segmentation conducted in the first stage. The work exploits the importance of spatial consistency amongst pixels, although their method is highly reliant upon the image segmentation stage. Jin et al. [21] focus on the colourisation of images containing natural objects using a reference colour image from which colour information is transferred. The authors require that the images are segmented and each segment consists of a single region that shares similar colour and texture statistics in nature. Colour is then transferred to pixels minimising a cost function measuring the consistency of colour in a neighbourhood and an intensity-to-colour correlation that is contained in a joint histogram. Xia [22] proposed a saliency guided approach in an attempt to align the colourisation to human perception. The approach first generates a saliency map of the reference and target images to predict the visual attention of human viewers, softly segmenting the images into foreground and background. Colour transfers are then made first to the foreground and then the background using a weighted colour transfer algorithm. Wu et al. [23] emphasise the use of high-level scene analysis to transfer colour between a colour image and grayscale target. The approach is heavily dependent upon the extraction of the foreground areas and the background from the images. Semantic correspondences between the regions in the reference and target images are then established with colour transferred between the corresponding regions. Gupta et al. [24] proposed a cascaded feature matching scheme to automatically find correspondences between superpixels of the reference and target images. Charpiat et al. [25] proposed an image colourisation method via multimodal predictions rather than choosing the most probable colour at the local level. Kuzovkin et al. [26] proposed a descriptor based image colourisation method and designed a novel regularisation scheme to smooth artefacts.

The proliferation of Internet images can be utilised for image colourisation. Liu et al. [27] recognised one of the causes of poor colourisation to be differences between a reference and destination image's illumination. Their method attempts to reduce these differences before the colour transfer process, reintroducing the destination illumination once the colour transfer process is complete. However, a major limitation is their need for multiple reference images in order to produce more reliable results, due to the problem of having both reflectance and illumination to solve. This is highlighted by the authors' illustrative examples, which are restricted to well known monuments and buildings for which multiple reference images are easily available. Chia et al. [28] propose an approach for semantic colourisation using Internet images. The user first provides segmentation clues for the major foreground objects in the image. The Internet is then searched for reference images based upon a semantic label given by the user and from the vast number of returned images a subset is chosen by means of a combined similarity metric. The method can produce multiple plausible colourised results for users to choose, although user effort is needed to assist in segmentation and label specification.

Deep learning has recently been applied to image colourisation [29]–[36]. Such methods are fully automatic, capable of colourising an input grayscale image without reference, although they need a very large training dataset and may fail to produce desired results when semantic ambiguities exist. Although recent work [7] can control the colourisation result by giving a reference colour image through minimising the errors between the colour distribution of the output image and that of the reference image, it is a global colour mapping method, and so may produce obvious artefacts since only the global colour distribution is used for guidance while the local texture features are ignored (see e.g. the second row of Fig. 1).

Some previous research considers colourisation of specific images, such as cartoons [37] and manga [38] which utilise the specific nature of these artistic forms. Some techniques exist that have different aims from colourising grayscale images but also bear some similarities, including colour transfer that transfers the colour styles from a reference colour image to a destination colour image (e.g. [39]), colour harmonisation that replaces colours in an image with a more aesthetically pleasing set of colours [40], and spot colour [41] where produced images are dominantly grayscale with typically compact regions containing colour. More knowledge about colourful image processing is reviewed in [42].

Our example-based automatic colourisation method is also

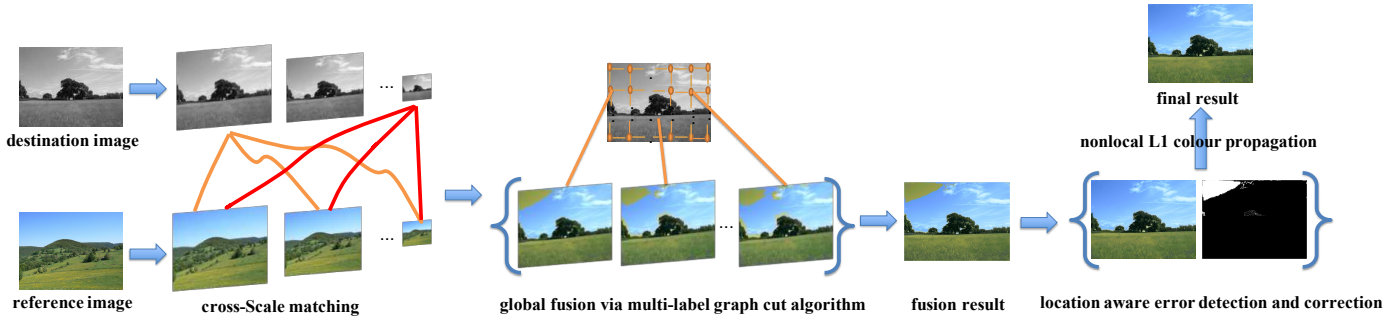

Fig. 2. The pipeline of the proposed method.

based on local texture matching. We introduce novel ideas of cross-scale matching and up-down colour location distribution to make colour matching more robust. We also develop a new confidence-weighted edge-preserving colour propagation method to produce robust colourisation with edges well preserved. As we will demonstrate later, our method produces plausible colourisation results even for challenging cases.

III. Proposed Algorithm

Our colourisation algorithm takes two images as input, a reference colour image, from which the colour information is transferred, and a destination grayscale image to which the colour information is transferred. The two images are expected to have a reasonably strong correlation in terms of texture content to obtain good colourisation. For each pixel in the destination image, we find a corresponding pixel in the reference image using cross-scale texture matching (Section A). The chrominance components of the matched pixels are transferred to the destination image to form a micro-scribble image. As only low-level texture information is used, the micro-scribble image may have significant mismatches semantically. In order to improve the matching result, up-down location statistics learnt from the reference image are used to detect and correct unreasonable matches. This simple strategy addresses a common class of semantic violation (Section B). In addition to the micro-scribble image, a normalised confidence map is also produced based on the texture matching results. Unlike previous work [3], [20] where the confidence map is thresholded to produce a sparse set of micro-scribbles, we propose to use confidence weighted optimisation that takes into account all the micro-scribbles with different weights (Section C). This not only uses more information from the matching but also avoids the task of choosing the threshold, for which an inappropriate value can substantially degrade the results. In addition, a nonlocal ℓ_1 propagation framework is proposed to achieve effective propagation while avoiding over-smoothing effect at edges. The framework of the proposed method is shown in Fig. 2.

A. Micro-scribbles using Cross-scale Texture Matching

Compared with the RGB colour space, the Lab colour model is designed to better approximate human vision, and is more convenient for colour editing [43], [44]. Therefore the

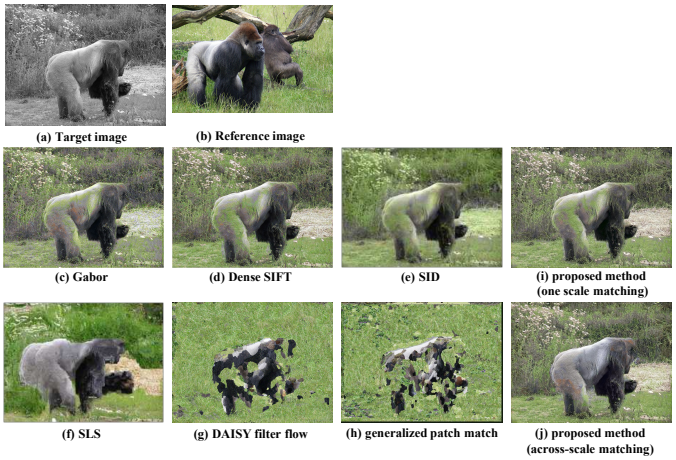

Fig. 3. Micro-scribble images by texture matching with different descriptors. After showing the grayscale destination image and the colour reference image, results with different feature matching methods are shown: (a) target image, (b) reference image, (c) Gabor feature, (d) Dense SIFT, (e) SID, (f) SiftFlow, (g) DaisyFlow, (h) generalized PatchMatch, (i) single scale Law's, (j) Law's with our cross-scale matching.

colourisation algorithm proposed in this paper is conducted in the Lab space. The reference colour image is first converted to Lab colour space where L is the luminance component and a and b are the two chrominance components. Then, we apply luminance normalisation for the reference image to decrease the global difference between the luminance values of the reference and destination images, as in [4]. The luminance of the destination grayscale image is maintained, and the chrominance components are transferred from the reference colour image independently.

Since the destination image only contains grayscale information, correspondence between the reference and destination images is established using the local texture information. However, the reference and destination images may have different scales (e.g. wide-field vs. close-up views) and different scales may also appear within the images (e.g. due to distance variation in a perspective view). In this paper, we propose to further use a cross-scale texture matching strategy to improve the correspondence for colourisation.

We first describe the texture descriptors for single level correspondence followed by generalisation to allow local textures to be matched across scales.

1) Local Texture Descriptors: The local texture descriptors need to be calculated on a large number of pixels. They should be distinctive but also efficient and stable. Welsh et al. [4] used luminance statistics with limited distinctiveness. Ying et al. [11] used co-occurrence matrix texture measures which while distinctive are expensive to compute. We instead use Law's texture measures which can be efficiently calculated via convolution in addition to some luminance statistics.

The Law's texture measures are based upon five convolution kernels each of length five although they are also available in lengths of three and seven [45]. The choice to use the kernels of length five was to provide an area around a pixel that was capable of providing enough detail to produce an accurate texture measure. The kernels used, shown below, are convolved together in pairs to produce twenty-five two dimensional convolution kernels.

$$\begin{aligned} L5 &= [1\ 4\ 6\ 4\ 1] \\ E5 &= [-1\ -2\ 0\ 2\ 1] \\ S5 &= [-1\ 0\ 2\ 0\ -1] \\ W5 &= [-1\ 2\ 0\ -1\ 1] \\ R5 &= [1\ -4\ 6\ -4\ 1] \end{aligned}$$

The L5L5 convolution kernel is used for normalisation and hence is not calculated for this particular application. The result of applying the remaining convolution kernels is a set of twenty-four kernels which then undergo a further convolution to replace each element with the average of the surrounding 15 $\times$ 15 neighbourhood. In addition to Law's texture measures, the luminance value as well as the average and standard deviation of the luminance value in the 7 $\times$ 7 neighbourhood are also included. This gives a 27 dimensional local texture descriptor for each sampled pixel.

Although there are many existing scale invariant or multiscale feature descriptors, such as SIFT [5], Gabor [46], SID [47], SLS [48], etc., most of them are not suitable for the task of image colourisation. Multiscale feature descriptors, such as Gabor, can produce features with different scales and directions simultaneously. However, they treat the features from different scales equally and do not match features of one scale in the reference image with the features of a different scale in the destination image, as shown in Fig. 3(c). SIFT feature is sparse and cannot be used to produce detailed matching for transferring colours. The dense descriptors such as dense SIFT (DSIFT), can produce a feature vector for each pixel, however, DSIFT is not scale invariant, which may fail to find good matching (see Figs. 3(d)). Compared with DSIFT, SID is scale invariant and produces better results (see Fig. 3(e)), however, numerous matching errors have also occurred, such as the colour of hair has been mismatched to the green colour of the grass. In addition, the computation complexity of dense matching using SID feature is high. For the image with size 480×640, the matching process costs 933.91s[1], and the matching performance is poor compared with the proposed method (as shown in Fig. 3(j)) which only costs 117.35s.

There are also some existing cross-scale matching methods, such as SIFT flow [49] used in [48], DAISY filter flow [50] and the generalised PatchMatch [51]. However, as shown in Figs. 3(f-h), most of the existing cross-scale matching methods fail to align objects under large displacements. As shown in Fig. 3(i), a single-scale Law's feature produces similar matching results as SID features, although being more efficient. It is thus more suitable in our cross-scale texture matching framework (see details below) which produces much improved matching results (Fig. 3(j)).

2) Cross-scale Texture Matching and Global Fusion: Since the reference and destination images may contain content at different or even varying scales (due to changes of distance between objects and the camera, for example), in order to make use of the spatial coherence, a novel cross-scale texture matching method is proposed in this section to improve the robustness and quality of the colourisation results. Firstly, both the reference and destination images are repeatedly downscaled by a factor of $\sqrt{2}$ in each dimension (i.e. with half the pixels) to form image pyramids. This repeats until the number of pixels in either dimension drops below a threshold that is defined as 75 in this paper. The local texture descriptors extracted from pixels of each scale of the reference image compose the search trees, which can be searched efficiently using the approximate nearest neighbour (ANN) tree searching algorithm [52].

Each of the scaled reference images is searched for the best match sample for each of the scaled destination images. The matching distance d_p for pixel p using single scale texture matching is defined as the Euclidean distance in the 27 dimensional local texture descriptor. For each pixel in the destination image, there will be $m \times n$ matching results (as illustrated in Fig. 2), where m and n are the numbers of scales for the destination and reference images, respectively.

The suitable scale to match the reference image with the destination image is spatially varying so a naive solution would simply choose the scale that gives the minimum matching cost for each pixel out of $m \times n$ scale combinations. This approach however is not robust because these matching costs can be fairly close and making individual decisions can be severely affected by slight numerical differences. We observe that the spatially varying scales should also be spatially coherent in the majority of cases, and scale changes are relative rare, e.g. at the boundary of different objects. We thus formulate cross-scale matching as a labelling problem. For each pixel p of the destination image T, the goal of the cross-scale matching is to find a best labelling function $f : f(p) \rightarrow L$, where $L = \{1, 2, \cdots, mn\}$ is the label set. Ideally labelling should satisfy the following two criteria. Firstly, the best matching scale should have as small as possible matching error in the feature space. Secondly, the neighbourhood pixels should have matching results in the same scale as much as possible. A solution should try to achieve a good balance of these two criteria, which will enhance the scale consistency while combining good matches in the feature space.

The cross-scale matching problem can be solved by min-

[1]According to the default setting of code provided by the authors of [47], a 1008 dimensional feature will be extracted for each pixel. For computation efficiency, we collect all of the features of both target image and reference image, and then reduce the dimension to 30 by using PCA.

imising the following energy function $E(f)$:

$$\min_f E(f) = \sum_{p\in P} D_p(f_p) + \sum_{(p,q)\in\mathcal{N}} V_{p,q}(f_p, f_q), \quad (1)$$

where P represents the set of pixels in the destination image, $\mathcal{N}$ is the neighbourhood system on pixels, where 4-connected neighbourhood is used in this paper.

The first term $\sum_{p\in P} D_p(f_p)$ is the data cost energy, which measures the penalty of assigning label f_p to pixel p. In this paper, $D_p(f_p)$ is defined as the matching distance d_p for the scale f_p. The second term $\sum_{(p,q)\in\mathcal{N}} V_{p,q}(f_p, f_q)$ is called the smoothness energy, which enhances spatial smoothness, or in this case scale consistency. It measures the cost of assigning the labels f_p and f_q to adjacent pixels p, q. In this paper, we assume that the neighbourhood pixels should have matching results in the same scale as much as possible. Based on the assumption, the smoothness energy is defined as follows

$$V_{p,q}(f_p, f_q) = ((s_p - t_p) - (s_q - t_q))^2,$$

where $s_p = \lfloor f_p/n \rfloor$ means the scale of destination image and $t_p = f_p \mod n$ corresponds to the scale of reference image. n is number of scales for the destination image. s_q and t_q are defined similarly for pixel q. $s_p - t_p$ and $s_q - t_q$ are scale differences between the reference and destination for pixels p and q. $V_{p,q} = 0$ if such scale differences are the same for p and q, which includes the case $f_p = f_q$. The optimisation problem (1) can be efficiently solved by the global multi-label graph cut algorithm [53]. An example is shown in Fig. 3 to demonstrate the effectiveness of cross-scale matching. (i) and (j) are the results obtained using single scale and cross scale matching respectively. We can see that, e.g. the hair of the gorilla is mistakenly matched to the green grass because the local textures of the hair and the grass look somewhat similar at the initial scale. By using cross-scale matching, most wrong initial matches are corrected. Note that cross-scale matching is much more powerful than multiscale matching, e.g. using multiscale Gabor descriptors (c), because a specific scale is chosen rather than combining multiple scales with equal weights, and the reference and destination images of different scales (including scaling up and scaling down) may be matched, which is not considered using multiscale matching.

B. Location aware correction of matching results

Our cross-scale matching makes colour matching more robust. However, since only low-level texture features are used, it cannot eliminate semantically unreasonable matching entirely. An example is shown in Fig. 4(e). The blue sky is matched to the green grass. A human observer can obviously see this as semantically implausible. However, learning generic semantic constraints requires a large number of training images and sophisticated learning. We find that for examples similar to Fig. 4(e) where semantically incorrect colourisation is related to up-down location violation (i.e. grass should not appear above the sky), and such location based "knowledge" can be automatically learnt from the single reference image — in this case green pixels should not appear above majority of blue pixels as this hardly happens in the reference image. As we will demonstrate later, such semantic violation is often produced by existing colourisation methods and this simple strategy works well to fix such challenging cases.

TABLE I
THE PROBABILITY p_{ij} OF EACH UP-DOWN DISTRIBUTION PROBABILITY FOR THE EXAMPLE IN FIG. 4 (A). IMPLAUSIBLE COLOUR PAIRS WITH $p_{ij} < \gamma$ ARE HIGHLIGHTED. γ IS CHOSEN AS 0.1.

c_i \ c_j	1	2	3	4
1	-	**0**	**0.0359**	0.3159
2	1.0000	-	0.9935	0.9998
3	0.9644	**0.0060**	-	0.8635
4	0.6852	**0.0002**	0.1356	-

We now describe our algorithm. Firstly, the pixels of the reference image are clustered by k-means clustering according to colour components (a and b channels). The number of clusters k can be automatically determined, e.g. by maximising the Bayesian Information Criterion (BIC) [54]. The colour labels are denoted by $C = \{c_1, c_2, \ldots, c_k\}$. An example is shown in Fig. 4(c) where $k = 4$. To identify the implausible up-down relationship, we measure the probability p_{ij} of pixels with one colour label c_i being above pixels of another colour label c_j. The probability p_{ij} can be naively computed by checking each pair of pixels with specific labels in the image, and working out the proportion with c_i being above c_j. However, the computational complexity for directly counting an $\tilde{m} \times \tilde{n}$ image will reach $O((\tilde{m} \times \tilde{n})^2)$, where $\tilde{m}$ and $\tilde{n}$ are the height and width of the image, which is very expensive.

In this paper, we propose a fast algorithm for computing the probability p_{ij}, which gives the exact solution. Our observation is that the up-down relationship is only affected by the row, so all the pixels in the same row can be grouped and their contributions to the probability can be worked out together, without enumerating individual pixel pairs. For each row, we compute the histogram of colour labels, $h^r = (h_1^r, h_2^r, \ldots, h_k^r)$, where h_j^r means the number of pixels with colour label c_j in the r-th row. The probability p_{ij} can be computed efficiently by

$$p_{ij} = \frac{\sum_{r=1}^{\tilde{m}-1} \sum_{r'>r} h_i^r h_j^{r'}}{\sum_{r=1}^{\tilde{m}-1} \sum_{r'=1}^{\tilde{m}} h_i^r h_j^{r'}}, \quad (2)$$

The computational complexity of our method is $O(\tilde{m}\tilde{n} + \tilde{n}^2)$, where $O(\tilde{m}\tilde{n})$ is the time for building the row-based histograms, and $O(\tilde{n}^2)$ for working out Eq. (2).

For the reference image as shown in Fig. 4(b), the obtained up-down probabilities for colour label pairs are shown in Table I. We can see that the values of p_{12}, p_{32} and p_{42} are almost zero, which implies that the colours of green grass and tree c_1, c_3 and c_4 rarely appear above the blue sky c_2. It is in line with the real colour distribution of the reference image. We consider an up-down distribution as unreasonable if $p_{ij} < \gamma$, where γ is a given threshold.

For the destination image, given the original matching result (Fig. 4(d)), the colour labels can be directly obtained since the chrominance channels are always transferred from pixels in the reference image (Fig. 4(e)). The statistics of up-down probability distribution of colour pairs p'_{ij} can also be

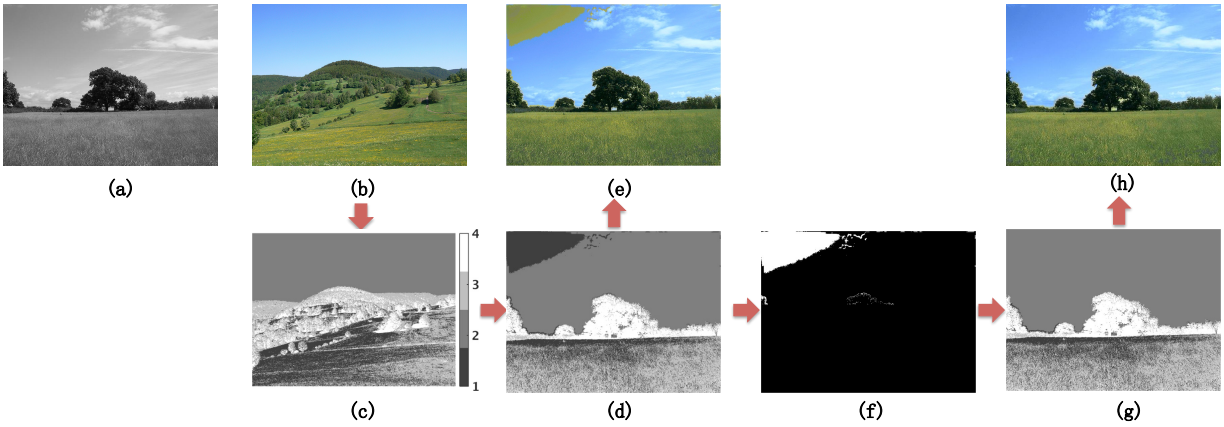

Fig. 4. The framework of the proposed location aware matching correction. (a) the reference image, (b) the destination image, (c) the cluster label of colours in the reference image, (d) the matching label in the destination image with semantic errors, (e) the colour matching result image, (f) the mismatched region identified using our location aware analysis, (g) the result after location aware matching correction, (h) the final colourised image with semantic errors fixed.

computed efficiently using the method (2). As we assume that the destination image has strong correlation with the reference image, if implausible up-down relationships appear in the matching result (i.e. if $p_{ij} < \gamma$ and $p'_{ij} > \gamma$), they are regarded as semantically wrong matches. For example, green appearing above blue as shown in Figure 4(e), rarely happens in the reference image. We assume the majority of pixels are correctly labelled, so we mark either those pixels with label c_i appearing above pixels with label c_j or pixels with label c_j appearing below those with label c_i as wrong pixels, depending on which group contains fewer pixels, as highlighted in Fig. 4(f).

When wrongly matched pixels are detected, we design an effective matching framework to correct them. Those pixels will be re-matched against reference pixels in the feature space, with identified implausible labels removed. In practice, this can be very efficiently implemented by building k ANN search trees, one for each colour cluster. We simply search for the best match among all the trees which are not excluded. For the example in Fig. 4, the matching results for the incorrect green region in Fig. 4(e) will be updated, with the green search tree excluded from matching. The best matched labels are then found (Fig. 4(g)), and finally the best matched colour with correct label is assigned to the query pixel (Fg. 4(h)), which produces plausible colourisation.

Fig. 5 illustrates the performance against several other popular example-based colourisation methods [14], [17], [18]. It is easy to find numerous matching errors in the results generated by these methods, e.g., the green colour is mapped above the blue sky in the first example, and the green colour is mismatched to the body of the pyramid while the blue colour is mapped to the grass, which is seldom found in the reference image in the second example. In contrast, these wrong regions are effectively corrected after the up-down location based correction, as shown in the last column of Fig. 5.

Our location-based correction is effective in identifying and fixing a class of semantic violations. A parameter γ is involved which is key to the performance. Fortunately, we find that the method is generally stable with changing γ. In Fig. 6, we vary γ massively from 0.0001 to 0.2, and the corresponding

Fig. 5. Examples of up-down location aware correction results. (a) the first column shows the reference and destination images. From the second column to the last column are the corresponding results of [14], [17], [18] and the proposed method.

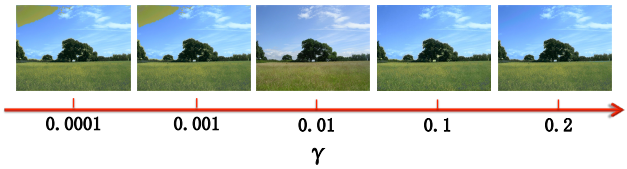

Fig. 6. The results of location aware correction using different values of γ.

result images are shown. We can see that when γ is too small, some mismatched regions may not be identified, and no semantic correction is performed. With the increasing value of γ, more and more pixels are corrected. $\gamma = 0.1$ is sufficient to produce correct results. The method however is not sensitive to the exact choice, as e.g. the same correct result is obtained with $\gamma = 0.2$. As it is safer to choose smaller γ to avoid overcorrection, the parameter γ is fixed as 0.1, which provides robust performance for all the experiments in this paper.

C. Confidence weighted nonlocal ℓ_1 colour propagation

Following Section A, after local texture matching, for each pixel p in the destination image we obtain the micro-scribble chrominance (Fig. 7(d)) and the matching distance d_p (Fig. 7(b)). The two chrominance channels a and b in the Lab colour space are treated separately for the transfer and the micro-scribble chrominance value for the channel being considered is denoted as $\lambda(p)$.

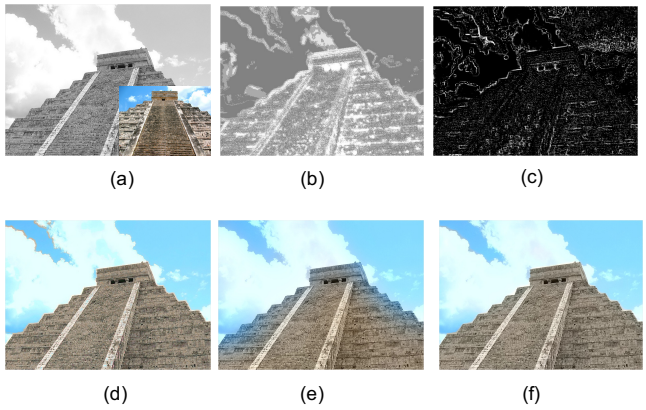

Fig. 7. Confidence weighted nonlocal ℓ_1 colour propagation. (a) the reference image and destination image. (b) the initial confidence map. (c) the normalised confidence map. (d) the initial colourisation result. (e) and (f) are respectively the colour propagation results of method [3] and our method.

To make propagation more robust, we do not treat each micro-scribble equally. Instead, a normalised confidence map is produced to be used as weights in the later propagation stage. For each pixel p, the confidence $w_c(p) \in (0, 1]$, where one is a high confidence and is calculated as shown in Eq. (3):

$$w_c(p) = \exp\left\{-\frac{d_p^2}{\tau^2}\right\}, \tag{3}$$

where τ is the scaling parameter. The value selected for τ is able to adjust the confident distribution. Instead of using a fixed parameter, we define τ based on the statistics of individual images to be colourised as in Eq. (4):

$$\tau^2 = -\frac{\mu^2}{\log c}, \tag{4}$$

μ is taken as the mean of the distances for all the micro-scribbles. To normalise the mapping for various images, Eq. (4) is defined such that the mean distance value should map onto a confidence level of c. In practice, to make the confident micro-scribbles more distinctive, a relatively small c is found to work well for various images; $c = 0.001$ is used in this paper.

Most of the existing colourisation methods use the confidence map to produce a sparse set of micro-scribbles, which involve a difficult task of choosing optimal threshold. In the propagation step, the least squares optimisation [3] is usually performed which results in obvious colour bleeding near edges, as shown in Fig. 7(e). In contrast, with our approach there is no need to adjust the parameter c for individual images and the same parameter is used for all the results in the paper. This adaptability is demonstrated also in Fig. 7(c), where the most confident pixels are highlighted.

We further propose a novel confidence weighted nonlocal ℓ_1 colour propagation method based on a dense micro-scribble image and a confidence map, used as a soft constraint in the optimisation framework. The proposed nonlocal ℓ_1 propagation framework suppresses over-smoothing at edges effectively.

For each pixel p, let $\boldsymbol{\alpha}(p)$ be the chrominance channel that is calculated (which can refer to the channel a or b), and $\boldsymbol{\lambda}(p)$ the micro-scribble value from the corresponding chrominance channel of the pixel p, the energy E is formulated as

$$E(\boldsymbol{\alpha}) = \|\nabla_w \boldsymbol{\alpha}\|_1 + \frac{\beta_1}{2}\sum_{p\in\Omega} w_c(p)\left(\boldsymbol{\alpha}(p) - \boldsymbol{\lambda}(p)\right)^2, \tag{5}$$

where $\|\nabla_w \boldsymbol{\alpha}\|_1 = \sum_{p\in\Omega}\sqrt{\sum_{q\in\Omega}(\boldsymbol{\alpha}(p) - w_{p,q}\boldsymbol{\alpha}(q))^2}$ is the nonlocal ℓ_1 regularisation term, which effectively enhances the smoothness while preserving edges and contrast of natural images well. Ω is the entire group of pixels. $w_{p,q}$ is the normalised nonlocal weight parameter defined as

$$w_{p,q} = \frac{1}{C_p}\exp\left\{-\frac{\|\boldsymbol{\alpha}(p+\cdot) - \boldsymbol{\alpha}(q+\cdot)\|^2}{2\sigma^2}\right\} \tag{6}$$

where $C_p = \sum_{q\in\Omega} w_{p,q}$ is the normalisation factor. "$+\cdot$" indexes pixels in the neighbourhood to form a vector. In this paper, the neighbourhood size is set to 5×5 and σ is fixed as 1. From the definition (6), the weight function is significant only if the patch around q has similar structure as the corresponding patch around pixel p. In order to improve the computational efficiency, for each pixel p, only 10 best neighbours are included by the semi-local searching within a window of 21×21 centred at p.

The second term $\sum_p w_c(p)(\boldsymbol{\alpha}(p) - \boldsymbol{\lambda}(p))^2$ is the data fidelity term weighted by the confidence map w_c. This term guarantees that pixels with high confidence tend to preserve micro-scribble values while other pixels with low confidence tend to receive propagation from corresponding high confidence pixels. β_1 is used to balance both terms, and we use $\beta_1 = 0.001$ for all examples.

The optimisation of problem (5) can be efficiently solved by the Split-Bregman algorithm [55]. In order to make the problem separable, a new variable $\mathbf{d}$ is introduced, and the original optimisation problem can be rewritten as follows

$$\min_{\mathbf{d},\boldsymbol{\alpha}} \|\mathbf{d}\|_1 + \frac{\beta_1}{2}\sum_{p\in\Omega} w_c(p)\left(\boldsymbol{\alpha}(p) - \boldsymbol{\lambda}(p)\right)^2, s.t.\ \mathbf{d} = \nabla_w \boldsymbol{\alpha}. \tag{7}$$

The hard constraint $\mathbf{d} = \nabla_w \boldsymbol{\alpha}$ can be guaranteed by the following efficient equivalent Bregman iteration approach,

$$\begin{aligned}(\boldsymbol{\alpha}^{k+1}, \mathbf{d}^{k+1}) &= \arg\min_{\boldsymbol{\alpha},\mathbf{d}} \|\mathbf{d}\|_1 \\ &+ \frac{\beta_1}{2}\sum_{p\in\Omega} w_c(p)\left(\boldsymbol{\alpha}(p) - \boldsymbol{\lambda}(p)\right)^2 + \frac{\beta_2}{2}\left\|\mathbf{d} - \nabla_w\boldsymbol{\alpha} - \boldsymbol{\Gamma}^k\right\|^2, \\ \boldsymbol{\Gamma}^{k+1} &= \boldsymbol{\Gamma}^k + \mathbf{d}^{k+1} - \nabla_w \boldsymbol{\alpha}^{k+1},\end{aligned} \tag{8}$$

where $\boldsymbol{\Gamma}$ is the Bregman variable. Then the optimisation problem (7) can be solved by splitting it into several easy subproblems, and each of the variables $\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\Gamma}$ can be updated in turn with other variables fixed.

$$\begin{aligned}&\boldsymbol{\alpha}^{k+1}: \text{Solve for } \boldsymbol{\alpha}: \\ &\beta_1\sum_{p\in\Omega} w_c(p)(\boldsymbol{\alpha}(p) - \boldsymbol{\lambda}(p)) - \beta_2 div_w(\mathbf{d}^k - \nabla_w\boldsymbol{\alpha} - \boldsymbol{\Gamma}^k) = 0. \\ &\mathbf{d}^{k+1}: \mathbf{d}^{k+1} = S_{\beta_2}(\nabla_w\boldsymbol{\alpha}^{k+1} + \boldsymbol{\Gamma}^k). \\ &\boldsymbol{\Gamma}^{k+1}: \boldsymbol{\Gamma}^{k+1} = \boldsymbol{\Gamma}^k + \mathbf{d}^{k+1} - \nabla_w\boldsymbol{\alpha}^{k+1}.\end{aligned} \tag{9}$$

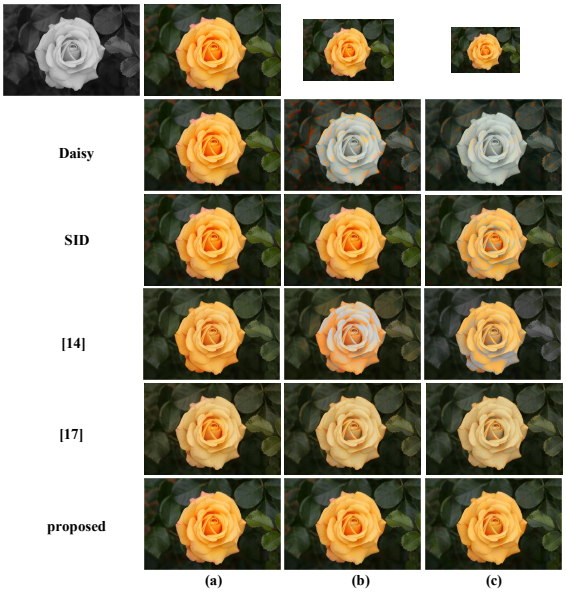

Fig. 8. Colourisation results with reference images of different scales.

where S_t is the soft-thresholding operator [55]. All of the parameters are set the same as proposed in the code[2].

Fig. 7 shows the propagation results of method [3] and the proposed method. It can be clearly seen that [3] (Fig. 7(e)) blurs the edge structure with the tone of entire image blended. In contrast, our method (Fig. 7(f)) not only suppresses the over-smoothing effect of boundaries, but also ensures the tone of image to be similar to the micro-scribble image.

IV. Experiments

In this section, we first design experiments to evaluate the robustness of the proposed algorithm to different scales of the reference image, followed by an ablation study to evaulate the effectiveness of key components. Then we compare the performance of the proposed algorithm by visual inspection against several state-of-the-art methods. Finally, a subjective user study is also performed to quantitatively evaluate the results of different methods.

We compare the colourisation results of the proposed method against seven state-of-the-art methods, including [7], [14], [17], [18], [32], [33], where [14], [17], [18] are example-based methods, and [7], [32], [33] are deep learning methods. For the example-based methods, the inputs are set as the same as ours (reference and destination image pairs). For the deep learning methods [7], [32], [33], only the destination images are provided. It is noted that method [7] can also cooperate with the reference image, so an extra experiment is to conduct [7] as an example-based colourisation method with the same input as our method. Therefore, 8 results will be provided for each example image. In order to guarantee fair comparison, the results of all of the algorithms are generated by the code provided by the authors.

Our method only involves a small number of parameters. As discussed, the method is insensitive to the exact choice of parameters so these are fixed in all our experiments. Specifically, the normalised parameter c in the confidence function (4) is set to 0.001, and the parameter γ in location correction is set to 0.1 in all of the experiments.

[2]http://www.math.sjtu.edu.cn/faculty/xqzhang/html/code.html

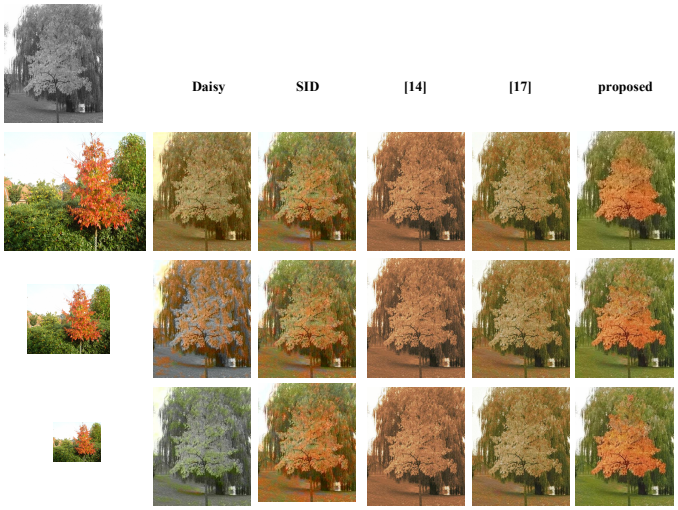

Fig. 9. Colourisation results with different scales of the reference images.

A. Robustness to the scale of reference image

Our cross-scale matching is able to handle cases where the reference and destination images are at different scales. We start with a simple case where the grayscale version of the reference colour image is used as the destination image, with the reference image at three scales (full size, 50% and 25%). We compare the colourisation results of the proposed algorithm against two state-of-the-art methods [14], [17]. According to the comparison in Fig. 3, the SLS, DAISY filter flow and the generalised patch match methods cannot produce meaningful results for the task of image colourisation. Therefore, in this experiment we only include the colourisation results by matching the DAISY feature [56] and SID feature [47] for comparison, which are known to be scale insensitive. When the reference and destination images are at the same scale, as shown in the second column, all the methods produce good results. However, when the reference image is downsampled, the performance of the other five compared methods suffer a sharp decline (the 3rd-4th columns). Compared with DAISY, SID is scale invariant and produces better results (the 3rd row). However, when the scale of the reference and destination images are far away from each other, the performance of using SID features declines (the 4th column). In contrast, the proposed algorithm is robust to different scales of reference images (bottom row). The cross-scale matching proposed in this paper automatically identifies suitable scales for matching and produces consistent colourisation results.

We perform further comparative experiments with reference images different from the destination images, also with varying scales. As shown in Fig. 9, our method produces consistent results, and so is robust to scale variation whereas the compared methods fail to produce good results when there are large scale differences between the reference and destination images.

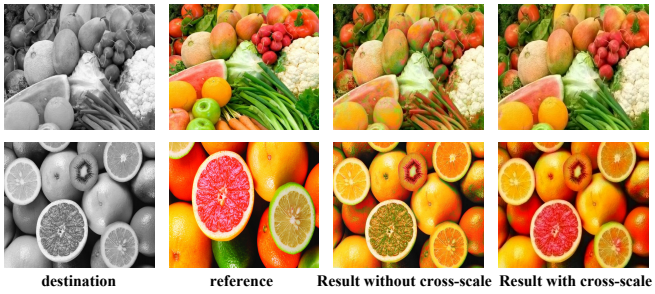

Fig. 10. Ablation study showing the effect of cross-scale matching.

B. Ablation Study

There are two main components critical to our final performance: cross-scale matching and up-down location aware correction. We perform an ablation study to assess the effectiveness of each component.

In the first experiment, cross-scale matching is disabled, and the single scale matching at the original scale of both destination and reference images is adopted. The experimental results are shown in Fig. 10. We can see that the radishes and spring onions are mismatched in the first example while the grapefruits get the wrong colour in the second example. Overall, most of the wrong matchings can be corrected by the cross-scale matching, as shown in the last column. As only low-level texture features are involved in the process of colourisation, it is difficult to distinguish different objects with similar textures, e.g., the intensity and texture of a lemon and a grapefruit are highly similar, and our method may occasionally generate mismatched results.

In the second experiment, the up-down location aware correction is disabled. From Fig. 11 we can see that many colour distributions which are seldom found in the reference images occur in the final results, such as the blue colour in the water and the red colour on the roof. By combining with the proposed up-down location aware correction, these mismatches can be effectively avoided.

C. Visual inspection and quantitative comparisons

In this section, the colourisation results of the proposed algorithm are evaluated by visual inspection. Fig. 14 shows a set of colourful natural images, which covers a wide variety of content including animals, fruits and landscapes.

We can see that in most cases the proposed algorithm (Fig. 14 (f)) achieves better results than the seven state-of-the-art methods in comparison. As we can see, due to texture similarity and scale variation, existing methods produce many mismatches for examples in the first three columns, resulting in semantically wrong results (red roof, blue water, etc.). With cross-scale matching and location aware correction, the proposed method produces plausible and semantically correct results. The examples in 4th - 9th columns contain complex textures of different scales, making colourisation difficult, as evidenced by substantial artefacts produced by alternative methods. The cross-scale matching proposed in this paper successfully finds the correct matching by choosing well-matched spatially coherent scales.

Method [14] is a global matching algorithm, which is based on finding and adjusting the zero-points of the histograms of both the reference and destination images. The global mechanism means the method can be less sensitive to local mismatches by confusing local textures. However, it results in mismatches in large regions when the zero-points based correspondence has error. In the worst case, the method does not reproduce the original colours from the reference images, but produces images with a colourless output for the example in the first column and generates uniform blended colour in the 5th, 8th and 9th columns (Fig. 14 (c)).

Method [17] solves the image colourisation problem by automatically selecting the best colour among a set of colour candidates via a total variational framework. In [17], strong regularisation coupling the channels of luminance and chrominance is proposed to preserve the image structures during colourisation, such as edges and colour consistency. However, the method does not take locality consistency into account in the process of choosing colour candidates, which leads to results with adjacent regions colourised with different tones (Fig. 14 (d)). Method [18] utilises spatial coherence around image structure by adopting an edge-aware texture descriptor based on region covariance, and reduces some misleading associations between reference and destination regions; however, the method may still produce some artefacts due to texture similarity and scale variation (Fig. 14 (e)).

Compared with example-based methods, the deep learning based colourisation algorithms [7], [32], [33] use millions of images for training the neural networks. In general, they can generate reasonable colour images, such as shown in Fig. 14 (g-j). However, they are distinct from the reference image, and these methods may still produce artefacts such as wrong colours (e.g. the blue reflection of a building in the second column for [32], [33], and the nearly uniform green colour output in the last column of Fig. 14 (i) for [7]).

Compared with methods [32], [33], [7] can not only produce an output without reference by the pretrained feed-forward network, but also cooperate with a reference image. Given a colour reference image, [7] produces a colourisation result by minimizing the errors between the prediction of colour distribution and the groundtruth colour distribution of the reference image. The colourisation results equipped with the reference images are shown in Fig. 14 (j). Compared with Fig. 14 (i), more flexible colours from the reference images are produced. However, the method [7] is a global colour mapping method, and many obvious artefacts can be found since only global colour distribution is used to guide the colourisation, while the local texture features are ignored, such as shown in the 5th - 7th columns in Fig. 14 (j).

D. Subjective user study

In addition to visual inspection, we would like to also make quantitative comparisons with existing methods. However, it is known that standard signal measures such as the standard Peak Signal to Noise Ratio (PSNR) can deviate substantially from human perceptual differences. Improved methods have been developed for image quality assessment in general, such

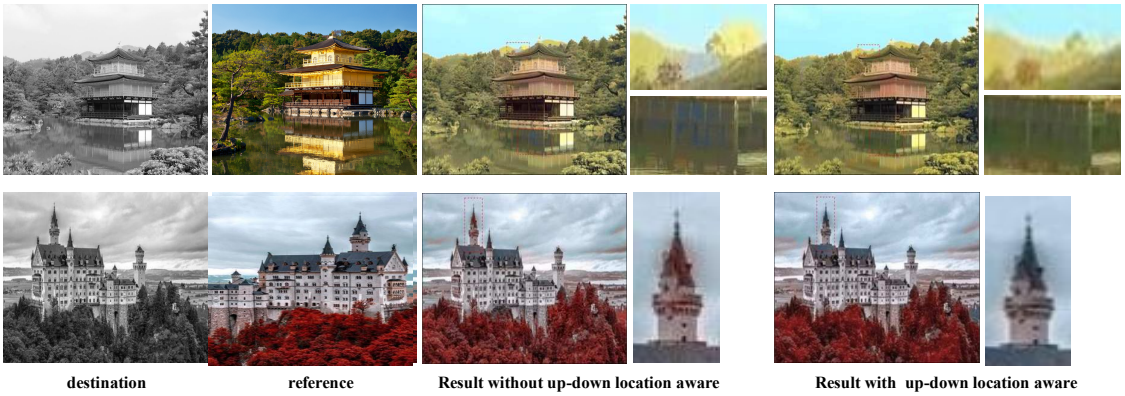

Fig. 11. Ablation study demonstrating the effect of up-down location aware correction.

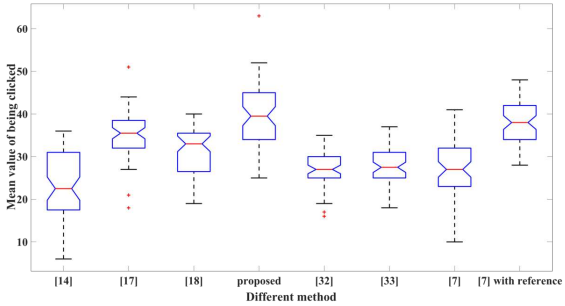

Fig. 12. The distribution of user preference for the first user study.

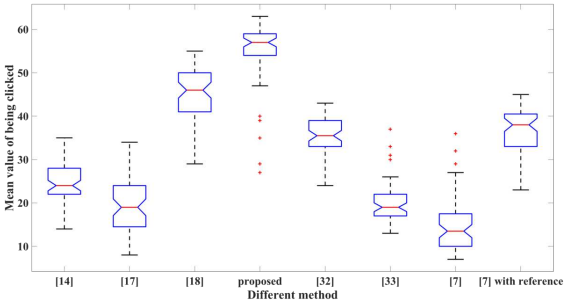

Fig. 13. The distribution of user preference for the second user study.

as Structural SIMilarity (SSIM) [57]. However, such measures are not appropriate for the task of image colourisation, e.g. because colourisation different from the ground truth may still be perfectly plausible. Therefore, in order to make a fair comparison, two user studies are designed to quantitatively evaluate our method against other seven methods.

The first user study evaluates the quality of the colourisation results, without considering reference images. The purpose of the first user study is to study whether the methods can produce plausible results. For each test image, every pairwise combination of results generated by different methods is shown and the user is asked to choose either of them according to their perceptual plausibility. In the second user study, in addition to pairwise results generated by different methods, the reference images used in example-based methods [14], [17], [18] and deep learning method [7] are also shown to the user. The user is asked to choose which result better matches the colour of the given reference image.

Each user study is designed using the 2AFC (Two-Alternative Forced Choice) paradigm, which is widely used in psychological studies, because it is simple and reliable. For meaningful comparisons with affordable user effort, we use all the examples in Fig. 14 containing 9 test images and their colourisation results generated by our method as well as 7 state-of-the-art methods. 60 users with age between 15 and 60 were invited to participate in the user study. The detailed results are given in the supplementary material. To avoid bias, we randomise the order of image pairs shown to the participants and their left/right positions. Altogether, results of each method are compared with $9 \times 7 = 63$ results of alternative methods. We record the total number of user preference (i.e. clicks) for each method, and treat them as random variables.

The distributions of user preference for each method of two user studies are shown in Figs. 12 and 13. The red line in these two box figures means the average score of each method. As results of each method are compared with other seven methods, the highest score is $9 \times 7 = 63$. From Fig. 12, we can see that deep learning methods can generate plausible results, especially the most recent method [7] gains the second highest score. In the second user study (Fig. 13), with the guidance of reference image, most of the users prefer the results of the proposed method. Compared with the results of the first user study, the scores of most deep learning methods reduce dramatically because these methods are not reference based so in many cases the results are not related to the reference images. We can see that [7] can get better performance when equipped with reference images.

In addition to the average scores, a one-way analysis of variance (ANOVA) is used for statistical analysis of the user study results. ANOVA is designed to determine whether there are any significant differences between the means of two or more independent (unrelated) groups. It returns the p-value for the null hypothesis that the means of the groups are equal. The smaller the p-value obtained by ANOVA, the higher chance that the null hypothesis is rejected, meaning the groups differ more significantly. In this paper, the p-values are computed between our method and each compared method, and the results are shown in Table II. We can see that in the first user

study, all of the p-values except for method [7] with reference image are small ($\ll 1e-3$) which implies that the judgements of all users on the other 6 methods are statistically significant. While our method has higher average score compared to [7], the difference is not statistically significant. In the second user study, all of the p-values are smaller than $1e-7$, which implies the judgements of all users are statistically significant, and combined with Fig. 13, we can see that the majority of the users prefer the method proposed in this paper which has the highest mean score.

The results of the subjective user study are in line with our expectations. Method [7] can produce plausible colourisation results without taking into account the reference as shown in the first user study, while the proposed method can produce the best example-based colourisation results, as shown in the second user study.

V. Conclusion and Future Work

In this paper, we propose a novel example-based image colourisation method with three major technical advances, namely cross-scale matching, location aware correction, and confidence weighted structure preserving colour propagation using nonlocal ℓ_1 optimisation. These have addressed major limitations of existing methods, and produce substantially improved results, as shown by extensive comparisons with state-of-the-art methods. Our method only involves a couple of insensitive parameters, which are fixed in all our experiments. Our location aware correction is able to learn statistics from the single reference image. As we have shown, it is capable of correcting semantic anomalies. However, this method relies on the assumptions of upright camera orientation, and semantic similarity between reference and destination images. While such assumptions are generally reasonable, they are not always true. In the future, we would like to learn semantics using a large number of training images and incorporate such prior knowledge in example-based colourisation, retaining the flexibility of controlling colour styles.

Acknowledgements

The work of Bo Li was supported by the Natural Science Foundation of China (NSFC) under Grant 61562062, Grant 61762064, Grant 61866027, and Jiangxi key R&D plan (20171BBE50013).

References

[1] A. A. Gooch, S. C. Olsen, J. Tumblin, and B. Gooch, "Color2gray: salience-preserving color removal," in *ACM Trans. Graph.*, vol. 24, no. 3, 2005, pp. 634–639.

[2] M. Grundland and N. A. Dodgson, "Decolorize: Fast, contrast enhancing, color to grayscale conversion," *Pattern Recognition*, vol. 40, no. 11, pp. 2891–2896, 2007.

[3] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.

[4] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[7] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 119, 2017.

[8] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1120–1129, 2006.

[9] D. Nie, Q. Ma, L. Ma, and S. Xiao, "Optimization based grayscale image colorization," *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1445–1451, 2007.

[10] A. Balinsky and N. Mohammad, "Colorization of natural images via L1 optimization," in *IEEE Workshop on Applications of Computer Vision*, 2009, pp. 1–6.

[11] J. Ying and L. Ji, "Pattern recognition based color transfer," in *IEEE Intl. Conf. on Computer Graphics, Imaging and Vision: New Trends*, 2005, pp. 55–60.

[12] F. J. Ferri, J. V. Albert, and E. Vidal, "Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 29, no. 5, pp. 667–672, 1999.

[13] T. Chen, Y. Wang, V. Schillings, and C. Meinel, "Grayscale image matting and colorization," in *Asian Conference on Computer Vision*, 2004, pp. 1164–1169.

[14] S. Liu and X. Zhang, "Automatic grayscale image colorization using histogram regression," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1673–1681, 2012.

[15] S. H. Kang and R. March, "Variational models for image colorization via chromaticity and brightness decomposition," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2251–2261, 2007.

[16] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 298–307, 2014.

[17] F. Pierre, J.-F. Aujol, A. Bugeau, N. Papadakis, and V.-T. Ta, "Luminance-chrominance model for image colorization," *SIAM J. Imaging Sciences*, vol. 8, no. 1, pp. 536–563, 2015.

[18] B. Arbelot, R. Vergne, T. Hurtut, and J. Thollot, "Automatic texture guided color transfer and colorization," in *Expressive*, 2016.

[19] B. Li, F. Zhao, Z. Su, X. Liang, Y. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5188–5202, 2017.

[20] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Eurographics Symp. on Rendering*, vol. 2, 2005.

[21] S.-Y. Jin, H.-J. Choi, and Y.-W. Tai, "A randomized algorithm for natural object colorization," in *Computer Graphics Forum*, vol. 33, no. 2, 2014, pp. 205–214.

[22] J. Xia, "Saliency-guided color transfer between images," in *Advances in Visual Computing*, 2013, pp. 468–475.

[23] F. Wu, W. Dong, Y. Kong, X. Mei, J.-C. Paul, and X. Zhang, "Content-based colour transfer," in *Computer Graphics Forum*, vol. 32, no. 1, 2013, pp. 190–203.

[24] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 369–378.

[25] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," *Computer Vision–ECCV 2008*, pp. 126–139, 2008.

[26] D. Kuzovkin, C. Chamaret, and T. Pouli, "Descriptor-based image colorization and regularization," in *International Workshop on Computational Color Imaging*. Springer, 2015, pp. 59–68.

[27] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, p. 152, 2008.

[28] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," *ACM Trans. Graph.*, vol. 30, no. 6, p. 156, 2011.

[29] Z. Cheng, Q. Yang, and B. Sheng, "Colorization using neural network ensemble," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5491–5505, 2017.

[30] A. Deshpande, J. Rock, and D. Forsyth, "Learning largescale automatic image colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 567–575.

[31] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 415–423.

TABLE II
THE P-VALUES OF ANOVA TEST COMPARING THE PROPOSED METHOD AGAINST OTHER METHODS.

method	[14]	[17]	[18]	[32]	[33]	[7]	[7] with reference
p-value (1st user study)	1.0727e-20	1.2069e-04	1.7829e-10	1.3525e-20	5.6336e-18	4.7435e-17	0.0673
p-value (2nd user study)	1.2464e-41	4.8198e-44	2.6079e-08	4.3135e-26	9.6833e-46	7.9269e-51	1.5931e-22

[32] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision – ECCV 2016: 14th European Conference*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, 2016, pp. 649–666.

[33] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 110, 2016.

[34] M. M. Larsson Gustav and G. Shakhnarovich, "Learning representations for automatic colorization," in *Computer Vision – ECCV 2016: 14th European Conference*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, 2016, pp. 577–593.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[36] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 47, 2018.

[37] D. Sỳkora, J. Buriánek, and J. Žára, "Unsupervised colorization of black-and-white cartoons," in *Symposium on Non-photorealistic animation and rendering*, 2004, pp. 121–127.

[38] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, 2006.

[39] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, no. 5, pp. 34–41, 2001.

[40] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 624–630, 2006.

[41] P. L. Rosin and Y.-K. Lai, "Non-photorealistic rendering with spot colour," in *Symposium on Computational Aesthetics*, 2013, pp. 67–75.

[42] A. Artusi, F. Banterle, T. Aydin, D. Panozzo, and O. Sorkine-Hornung, *Image Content Retargeting: Maintaining Color, Tone, and Spatial Consistency*, 1st ed. CRC Press, September 2016.

[43] A. Khandual, G. Baciu, and N. Rout, "Colorimetric processing of digital colour image!" *International Journal*, vol. 3, no. 7, 2013.

[44] S. Süsstrunk, R. Buckley, and S. Swen, "Standard RGB color spaces," in *Color and Imaging Conference*, vol. 1999, no. 1. Society for Imaging Science and Technology, 1999, pp. 127–134.

[45] K. I. Laws, "Rapid texture identification," in *24th annual technical symposium*. International Society for Optics and Photonics, 1980, pp. 376–381.

[46] J. G. Daugman, "Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[47] I. Kokkinos, M. Bronstein, and A. Yuille, "Dense scale invariant descriptors for images and surfaces," Ph.D. dissertation, INRIA, 2012.

[48] T. Hassner, V. Mayzels, and L. Zelnik-Manor, "On SIFTS and their scales," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1522–1528.

[49] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT flow: Dense correspondence across different scenes," in *European conference on computer vision*. Springer, 2008, pp. 28–42.

[50] H. Yang, W.-Y. Lin, and J. Lu, "Daisy filter flow: A generalized discrete approach to dense correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3406–3413.

[51] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *European Conference on Computer Vision*. Springer, 2010, pp. 29–43.

[52] D. M. Mount and S. Arya, "ANN: library for approximate nearest neighbour searching," 2010.

[53] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[54] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *International Conference on Machine Learning*, 2000, pp. 727–734.

[55] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imaging Sciences*, vol. 3, no. 3, pp. 253–276, 2010.

[56] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 815–830, 2010.

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

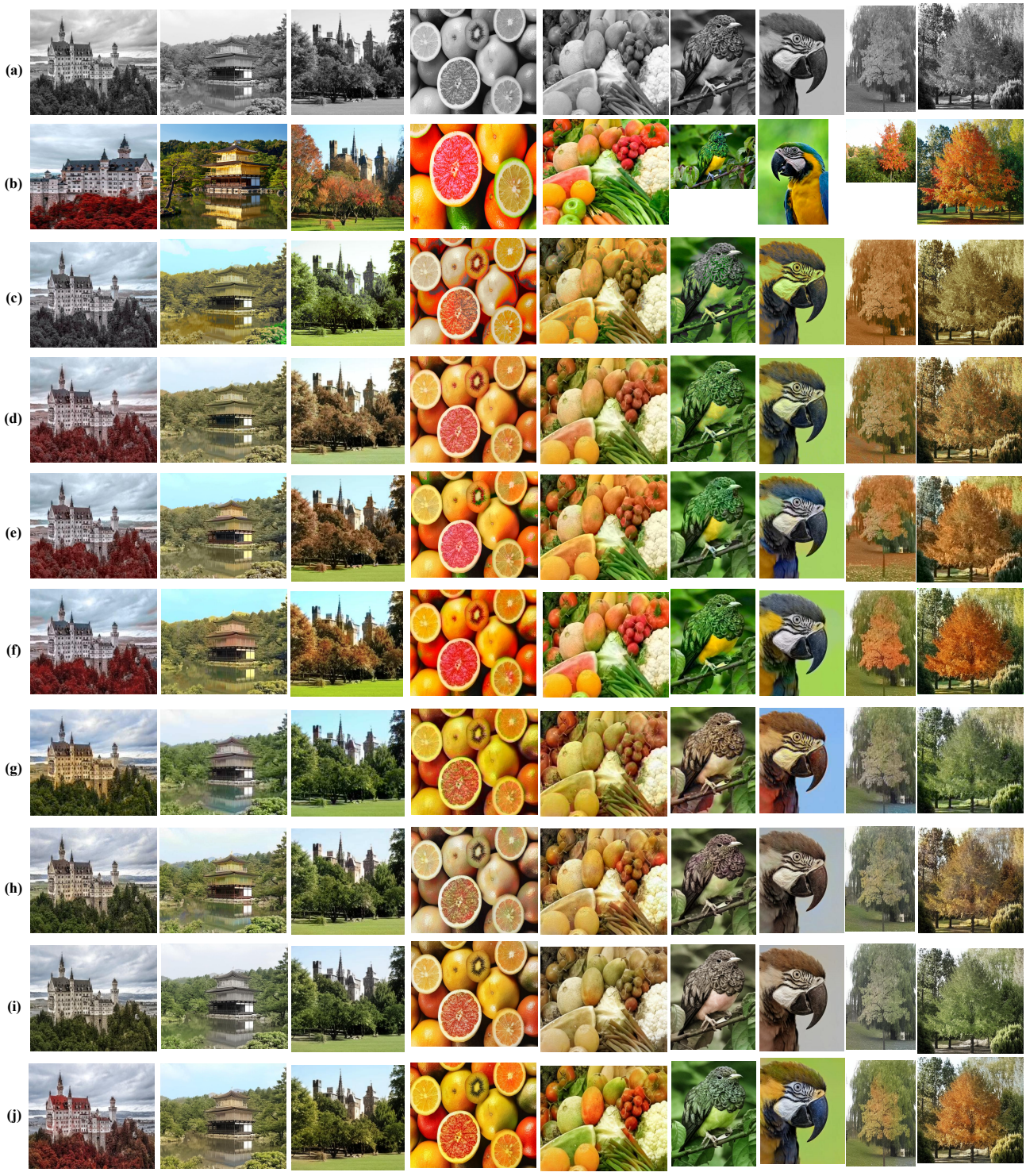

Fig. 14. Comparison of our colourisation results with alternative methods. (a) destination gray image, (b) reference image, (c) method [14], (d) method [17], (e) method [18], (f) proposed method, (g) method [32], (h) method [33], (i) method [7] without reference and (j) method [7] with reference.